Data collection and quality challenges in deep learning: a data-centric AI perspective

Year: 2021 | Citations: 440 | Authors: Steven Euijong Whang, Yuji Roh, Hwanjun Song, Jae-Gil Lee

---

Abstract

Data-centric AI is at the center of a fundamental shift in software engineering where machine learning becomes the new software, powered by big data and computing infrastructure. Here, software engineering needs to be re-thought where data become a first-class citizen on par with code. One striking observation is that a significant portion of the machine learning process is spent on data preparation. Without good data, even the best machine learning algorithms cannot perform well. As a result, data-centric AI practices are now becoming mainstream. Unfortunately, many datasets in the real world are small, dirty, biased, and even poisoned. In this survey, we study the research landscape for data collection and data quality primarily for deep learning applications. Data collection is important because there is lesser need for feature engineering for recent deep learning approaches, but instead more need for large amounts of data. For data quality, we study data validation, cleaning, and integration techniques. Even if the data cannot be fully cleaned, we can still cope with imperfect data during model training using robust model training techniques. In addition, while bias and fairness have been less studied in traditional data management research, these issues become essential topics in modern machine learning applications. We thus study fairness measures and unfairness mitigation techniques that can be applied before, during, or after model training. We believe that the data management community is well poised to solve these problems.